PrivateBus: Privacy Identification and Protection in Large-Scale Bus WiFi Systems

ZHIHAN FANG, Rutgers University, USA
BOYANG FU, Rutgers University, USA
ZHOU QIN, Rutgers University, USA
FAN ZHANG, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd.
DESHENG ZHANG, Rutgers University, USA

Recently, the ubiquity of mobile devices leads to an increasing demand of public network services, e.g., WiFi hot spots. As a part of this trend, modern transportation systems are equipped with public WiFi devices to provide Internet access for passengers as people spend a large amount of time on public transportation in their daily life. However, one of the key issues in public WiFi spots is the privacy concern due to its open access nature. Existing works either studied location privacy risk in human traces or privacy leakage in private networks such as cellular networks based on the data from cellular carriers. To the best of our knowledge, none of these work has been focused on bus WiFi privacy based on large-scale real-world data. In this paper, to explore the privacy risk in bus WiFi systems, we focus on two key questions *how likely bus WiFi users can be uniquely re-identified if partial usage information is leaked* and *how we can protect users from the leaked information.* To understand the above questions, we conduct a case study in a large-scale bus WiFi system, which contains 20 million connection records and 78 million location records from 770 thousand bus WiFi users during a two-month period. Technically, we design two models for our uniqueness analyses and protection, i.e., a *PB-FIND* model to identify the probability a user can be uniquely re-identified from leaked information; a *PB-HIDE* model to protect users from potentially leaked information. Specifically, we systematically measure the user uniqueness on users' finger traces (i.e., connection URL and domain), foot traces (i.e., locations), and hybrid traces (i.e., both finger and foot traces). Our measurement results reveal (i) 97.8% users can be uniquely re-identified by 4 random domain records of their finger traces and 96.2% users can be uniquely re-identified by 5 random locations on buses; (ii) 98.1% users can be uniquely re-identified by only 2 random records if both their connection records and locations are leaked to attackers. Moreover, the evaluation results show our *PB-HIDE* algorithm protects more than 95% users from the potentially leaked information by inserting only 1.5% synthetic records in the original dataset to preserve their data utility.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; **Privacy protections**; • **Human-centered computing →** *Ubiquitous and mobile computing*;

Additional Key Words and Phrases: uniqueness, bus WiFi, finger traces, foot traces, hybrid traces

ACM Reference Format:
Zhihan Fang, Boyang Fu, Zhou Qin, Fan Zhang, and Desheng Zhang. 2020. PrivateBus: Privacy Identification and Protection in Large-Scale Bus WiFi Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 9 (March 2020), 23 pages. https://doi.org/10.1145/3380990

Authors' addresses: Zhihan Fang, Rutgers University, Piscataway, NJ, 08854, USA, zhihan.fang@cs.rutgers.edu; Boyang Fu, Rutgers University, USA; Zhou Qin, Rutgers University, USA; Fan Zhang, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd. Desheng Zhang, Rutgers University, USA, desheng.zhang@cs.rutgers.edu.

2474-9567/2020/3-ART9 $15.00
https://doi.org/10.1145/3380990

1 INTRODUCTION

We are living in a digital era and experiencing a rapid growth of Internet access to public networks. The number of public WiFi APs (Access Points) deployed has reached 94 million globally and is expected to grow to 549 million in 2022 [10]. This growth is motivated by the prevalence of portable computing devices, e.g., there are about 3 billion smartphone users in 2018 in the world. As a promising technology, public transport WiFi enables high connection speeds and is a cost-effective way for passengers to access Internet services with their portable devices on public transportation systems such as buses and subways. Besides, it promotes potential commercial values for service providers, e.g., attracting more passengers and customizing advertisements for passengers. Because of those benefits, large cities such as New York City [18], Los Angeles [8] and Shenzhen [14] [17] [41] [34] have equipped their bus systems with WiFi services, which cover around 100% bus passengers in those cities.

Even though bus WiFi services provide passengers with highly convenient and ubiquitous Internet access, there is a growing concern about privacy leakage among bus WiFi users due to its open environment [7] [23]. Different from the WiFi hot spots at home or business environments, which are always protected with techniques such as WEP, WPA, WPA2 encryption [22] [3], the public transport WiFi is difficult to deploy with these protection techniques since it needs to accommodate on-the-go passengers. As a result, most bus WiFi hot spots are public without a security guarantee. Furthermore, bus WiFi operators have been collecting detailed bus WiFi usage records. Those records can be shared with other third parties such as advertisers for commercial purposes or researchers for research purposes. On the other hand, the usage records from real-world users are possible to be leaked due to the public environment on a bus and open nature of bus WiFi systems. For example, someone can remove WiFi devices in buses for vandalism or theft. With the leaked information, there is a potential risk that a user can be uniquely re-identified in bus WiFi systems. What is worse, most users are neglecting the privacy threats because there is no systematical work to quantify the privacy leakage in a public transport WiFi system. For example, if an attacker has access to user records of a bus WiFi system, e.g., a developer from advertisement company, he/she can get on that bus every day and saw who was using phone for what purposes (e.g., a girl was checking email at 10am), and then he/she can re-identify this real-world user from records later.

Researchers have accumulated abundant knowledge for user uniqueness and privacy from many large-scale systems such as location uniqueness [5], smartphone applications [36] [31] and passenger re-identification from their historical trajectories [30] [37] [40] [33]. Those works are built on systems with stationary sensing where stationary sensors are deployed to sensing locations, e.g., cell towers are stationary, regular WiFi APs are stationary, payment devices are stationary. Therefore, those studies cannot capture the correlation between continuous mobility patterns of passengers and online activities. Moreover, previous works mainly focus on the general setting of human activities, which is difficult to narrow down to the uniqueness in public transportation and their correlation with commuting patterns of passengers.

To the best of our knowledge, little work, if any, has been focused on uniqueness analyses and protection in large-scale mobile bus WiFi sensing systems or mobile WiFi systems in a generic setting. This is mainly because it is challenging to access large-scale bus WiFi usage data. Bus WiFi sensing systems are different from other works from three perspectives: (i) A bus WiFi system consists of mobile access points moving with passengers, in contrast to other systems such as cellular networks with stationary towers. Thus users will be mostly connected to a single access point even with long-distance travel whereas users are mostly connected to different cellular towers and stationary access points in cellular networks or stationary WiFi; (ii) In a bus WiFi setting, user devices and access points are close to each other in buses. Typically, real-time locations of buses are public information, e.g., for bus arrival prediction services. It leads to fine-grained and continuous location information on users' Internet access records, in contrast to discrete coarse-grained locations (tower level) of cellular networks. Based on those unique properties of bus WiFi systems, we conduct our analyses to (i) compare fine-grained hybrid traces, i.e. foot traces (i.e., offline physical locations) and finger traces (i.e., online visiting behaviors), with single

types of traces, which are widely investigated in other studies; (ii) study user privacy with users' commuting patterns such as commuters and visitors. Even though cellular networks can collect locations of users' connected towers, the sensing granularity is coarse-grained on both spatial and temporal dimensions. Besides, it is observed that users are not always connected to the nearest tower due to load balancing [28] [13] and Ping-Pong effect in cellular networks [19] [27]. Instead, mobile access points record locations of users passively and continuously with high accuracy. Therefore, bus WiFi systems record users' commuting patterns with high accuracy. Even though bus WiFi service providers and advertisers collected large-scale bus WiFi usage records to improve system efficiency and to customize advertisements based on user interests, these data are exclusively used for commercial purposes with extensive sharing. Recently, in the vision of smart cities, large-scale bus WiFi data have been shared with or traded to researchers and city planners for social good. Thus, the privacy concern increases when these data are centrally collected and copied for various purposes. Therefore, it is of great importance to identify and protect users' privacy in bus WiFi systems.

In this paper, we study the privacy identification and protection in bus WiFi systems from the *uniqueness* and *re-identification* perspective. More specifically, we study two key scientific questions that *how likely users can be uniquely re-identified from bus WiFi systems if partial traces (i.e., visiting records) are leaked* and more importantly *how we can protect users from the potentially leaked information.* To address the above questions, we conduct a case study based on a large-scale bus WiFi system in China. We collaborate with a leading bus WiFi service provider in China and are fortunate to access a large-scale bus WiFi dataset including 20 million connection records (which are defined as finger traces) and 78 million location records (which are defined as foot traces) for 770 thousand users during a two-month period. We summarize our contribution as follows:

- We utilize a real-world bus WiFi system and its connection records to understand and protect user uniqueness. Our study is based on the large scale privacy investigations covering 770 thousand users with a fairly complete data set including both finger traces (e.g., online visiting behaviors) and foot traces (e.g., offline physical locations) of users, which enable us to advance the state-of-the-art method for uniqueness analyses.
- We design a uniqueness analysis and protection framework named *PrivateBus*, which includes two key components, i.e., a uniqueness analysis model *PB-FIND* and a uniqueness protection model *PB-HIDE*. *PB-FIND* calculates the uniqueness score with leaked information under a low computing cost. *PB-HIDE* protects users from leaked information by inserting a limited amount of synthetic records in the original data.
- We implement and evaluate *PrivateBus* based on a large-scale bus WiFi system including 41 million connection and location records from 770 thousand bus WiFi users. Our privacy identification results reveal (i) 97.8% users can be uniquely re-identified by 4 random domain records of their finger traces; (ii) 96.2% users can be uniquely re-identified by 5 random locations on buses. (iii) 98.1% users can be uniquely re-identified by 2 random records if both their connection records and locations are leaked to attackers. Our privacy protection evaluation results reveal *PB-HIDE* algorithm protects more than 95% of users from potentially leaked information by inserting only 1.5% synthetic records in the original dataset to ensure the original dataset inserted with synthetic records is still useful when released, i.e., to protect the data utility while preserving the privacy. We will share our three days of data for 1,000 users as samples for the benefit of IMWUT community.

The remainder of our paper is organized as follows. Section 2 introduces the related work and section 3 describes datasets. We motivate our work in section 4 and present a uniqueness analysis model in section 5. Section 6 elaborates our uniqueness analysis results. Section 7 is a privacy protection model. Section 8 summarizes lesson learned and discussion, followed by the conclusion of our work in section 9.

2 RELATED WORK

Understanding privacy risk and crowd uniqueness in bus WiFi systems is of great importance for privacy protection and generalization of bus WiFi systems. In fact, there is a trend showing that users use their portable devices for Internet access more often than other activities. Investigating privacy risk on those portable devices has received considerable attention recently due to data availability. In table 1, we summarize existing studies based on a two-dimension taxonomy from both users' and service providers' perspectives: (i) user traces, i.e., single traces or hybrid traces. (ii) service end, i.e., users are connected to either stationary access points such as cellular towers or mobile access points such as mobile WiFi routers.

Table 1. Sensing Privacy Study Survey

Categories		Sensing End	
		Stationary Access Points	Mobile Access Points
User End	Single	[29] [42] [39] [4] [16] [31] [5] [7] [11] [20] [21] [15]	[24] [12]
	Hybrid	[31] [1] [38] [32]	*PrivateBus*

2.1 Stationary Access Points

Single types of traces: Recently, as the increase of mobile devices such as smart phones and tablets, high-speed and inexpensive Internet access are becoming crucial in daily life for most people. With the increase of Internet access demand, many infrastructures are built to provide data access services. Most of the studies of privacy on Internet users are built upon stationary sensors such as cellular towers in cellular networks [29] [42] [39], stationary WiFi access points [7], PoI locations [5]. The connected sensing devices change with human mobility and usage patterns. Most of the stationary sensing studies analyzed privacy risk in a single type of traces, i.e. foot traces or finger traces. For example, Shklovski *et al.* studied location tracking concerns from cellphone users on cellphone applications. Kandappu *et al.* developed a privacy-preserving mobile crowd-sourcing platform with a mobile client and a backend server [21]. Studies on human mobility revisitation revealed revisitation patterns of cellphone users in a city [20] [4] [11] [16] [15].

Hybrid types of user devices: Other studies are focused on uniqueness in more general settings and not restricted to a specific type of user traces. For example, Tu *et al.* studied the uniqueness in cellphone applications in cellular networks where the sensing end is the stationary cellular towers with finger traces and sparse observations of foot traces [31]. Almuhimedi *et al.* conducted a study to remind users of the privacy risk by installing a permission application on users' smart devices [1]. Xu *et al* revealed the privacy risk on aggregated level based on large-scale mobility data in cellular networks [38]. Tu *et al.* designed a synthetic model to generate human trajectories to protect user privacy and reserve data utilities [32].

2.2 Mobile Access Points

Single types of human traces: Compared with stationary access points, mobile access point operators provide services to user devices but passively collect detailed human mobility information based on access point locations. Users' locations are *passively* collected by the same access point when users are moving anytime when a user is connected to the system. Especially, such sensors show a high correlation with human commuting patterns. An onboard GPS is a mobile access point that can record locations of users accessing the vehicle. Mohamed *et al.* proposed a mobility privacy protection algorithm on small scale taxis in Bejing [24]. Douriez *et al.* studied the privacy protection among taxis in New York City [12].

Hybrid types of human traces (our work): To the best of our knowledge, little work, if any, has been conducted to analyze uniqueness and privacy risk in large scale bus WiFi systems, which track hybrid user traces continuously

with mobile access points. The privacy risk becomes a major concern in public WiFi services for users. Bus WiFi systems are unique with both finger traces and foot traces when users access to services. Our work is based on usage records of a large-scale bus WiFi system covering 770 thousand users in two months. Due to the open nature of public WiFi and the mobility nature of buses, public WiFi systems record both detailed finger traces and foot traces of users, which make our work different from others.

3 DATASET

In this section, we introduce a bus WiFi system and the data format collected in the system. Then we conduct a preliminary analysis based on the dataset.

3.1 Bus WiFi System

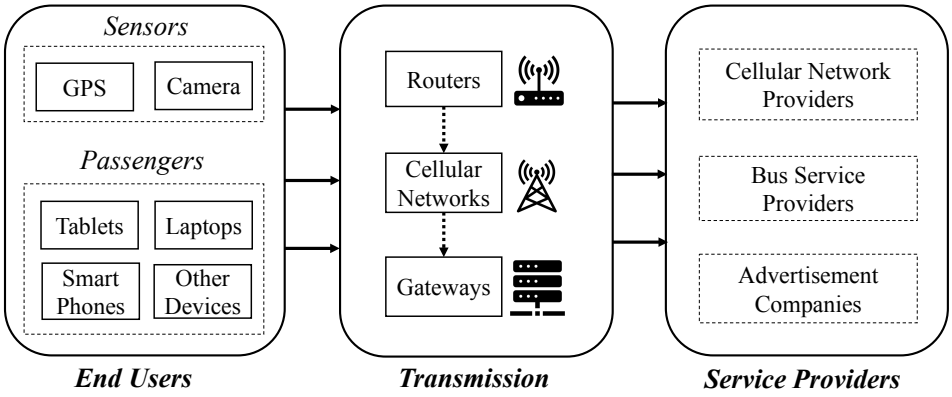


Fig. 1. Data Flow in Bus WiFi Systems

We collaborate with a bus operating company in China providing large-scale WiFi services. The workflow of the bus WiFi system is presented in Figure 1, which can be divided into three stages: (i) connections from end users' devices; (ii) data transmission from WiFi routers to gateways of cellular networks; (iii) data collection and analyses in service providers. In the first stage of the bus WiFi data flow, end users connect to WiFi routers on buses to access the Internet. There are two categories of data transmission relying on the bus WiFi for Internet access, i.e., data transmission for bus passengers and data transmission for bus onboard sensors. On one hand, bus passengers connect to the bus WiFi system with their portable devices such as smartphones, tablets, and laptops. On the other hand, data generated by onboard sensors, e.g., onboard GPS, videos from monitoring cameras, are uploaded to service providers with the bus WiFi system. In the second stage, the bus WiFi routers are connected to the nearby cellular towers. Both data transmission from passenger devices and onboard sensors requests utilities in cellular networks, which provide Internet services for the bus WiFi system. In the third stage, log records and onboard sensing data are collected by service providers including cellular network service providers, bus service providers and advertisement companies.

3.2 Data Format

We introduce our dataset collected by a large-scale bus WiFi system in Figure 2. The data are collected in two months from February to April in 2017 with 770 thousand users and 4,408 bus WiFi devices. Most of the bus WiFi devices (more than 4000) are deployed in Chinese City Shenzhen and the rest of them are deployed in other cities. The total data size is 40.7 GB. Based on different end devices, i.e. data from portable devices and data from

onboard GPS sensors, we define two types of traces: **finger traces** are records of passengers when they use bus WiFi for Internet connections; **foot traces** are GPS records of buses, which contain locations of bus WiFi users.

Due to the open nature of bus WiFi systems, it has a higher chance to reveal both the finger traces and foot traces of users. Compared with other systems such as (i) cellular networks, which record finger traces of users, and (ii) transportation systems, which record foot traces of passengers, both traces are collected by bus WiFi systems when users are connected.

Bus WiFI System			
Time Period	# of Users	# of Buses	Data Size
2017-02-02 to 2017-04-02	770 K	4408	40.7 GB
Finger Traces		***Foot Traces***	
# of Records	20 M	# of Records	78 M
Total Taffic	1.5 PB	Time Inteval	5 minutes
Attributes		***Attributes***	
Start Time	*End Time*		
URL	*Data Type*	*Longtitude*	*Latitude*
Traffic	*mac address*	*WiFi Device ID*	*Time*
Device Type	*WiFi Device ID*		

Fig. 2. Data Description

4 MOTIVATION

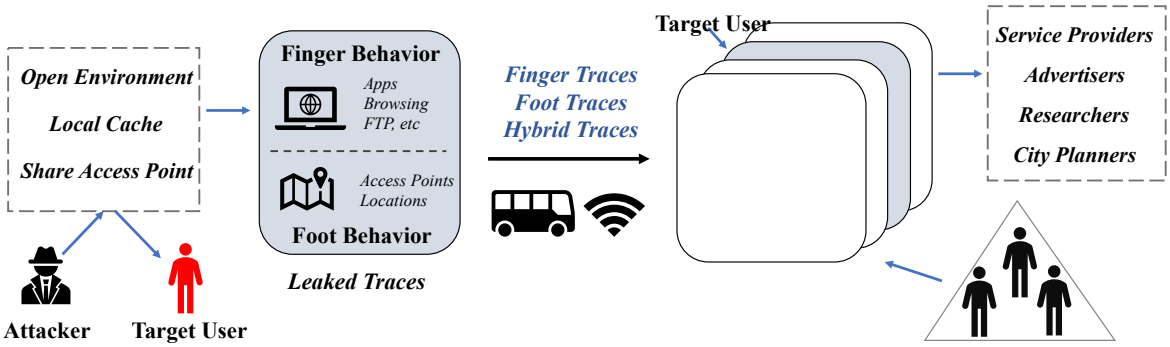


Fig. 3. Privacy Concern in Bus WiFi Systems

4.1 Privacy Concern

The service providers collect large-amount data for billing or user profiling purposes, e.g., potential advertisement. However, there is a potential risk of data leakage when data are centrally collected by many service providers [6][2]. Furthermore, since both onboard sensors and passenger devices rely on bus WiFi systems for data transmission, both finger traces, i.e., user connection records, and foot traces, i.e., user locations inferred from onboard sensor records, are recorded and collected. Due to the public nature of bus WiFi systems, there is a high risk of privacy breach for individual users. We illustrate a potential privacy leakage process in Figure 3. On one hand, service

providers collect historical records in the bus WiFi systems including both finger traces and foot traces for billing or profiling purposes. For instance, bus service operators and cellular network providers analyze the collected records for better caching strategies, e.g., caching popular videos to save cellular traffic. Advertisement companies model user interests from finger traces to customize advertisements for users. Moreover, the data are traded to city planners and researchers for social good. One the other hand, partial information of users are potentially leaked to attackers by observations, from local caches, and by the shared local network on the bus. As a result, understanding and modeling the privacy risk based on real-world data is crucial for both bus WiFi users and service providers in order to provide some privacy protection mechanisms. Specifically, we are interested in the possibility that a user can be uniquely re-identified from bus WiFi users with partial real-world observations.

4.2 Unique Patterns of Bus WiFi Users

Previous works have focused on uniqueness in general settings such as smartphone apps [31] and location-based services in public dataset [9] [5]. These studies reveal potential privacy risks in either finger traces or foot traces in smart phone apps and public map data. Compared with these studies, bus WiFi users show unique trace patterns in terms of both finger traces and foot traces. Figure 4 compares the user demands in a cellular network in Chinese city Shenzhen (we omit the data details due to space limitation) and the bus system in Shenzhen. *CDR* stands for call detail records and *Data* is the cellular data connection records of regular data services. This figure indicates the demand in Bus WiFi systems is more regular and concentrated than cellular networks, where the peak demands for Bus WiFi are between 7am-9am and 4pm-6pm, which are two peak hours commuting between home and work locations. We further study the entropy of bus passengers and cellular users with respect to foot traces and finger traces. The studied entropy is defined in Equation 1 where $X(s_j, u)$ is a function to count the frequency that a user u is observed at s_i, which is a location in foot traces and a connection domain in finger traces.

$$H(u) = -\sum_{j=1}^{n} p(X(s_j, u))log_2 p(X(s_j, u)) \tag{1}$$

Compared with cellular networks, bus systems have a high frequency on certain locations, e.g., bus routes commuting between home and work, and thus a low entropy on the observed locations. Figure 5 demonstrates that bus WiFi users show a more regular pattern on Internet access. For instance, we observed that few users use online payments in buses due to privacy concerns. In conclusion, it is easier to analyze the usage patterns and potentially expose the identities of bus WiFi users than cellular users.

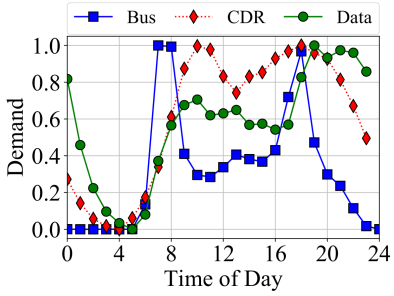


Fig. 4. Bus and Cellular Demand

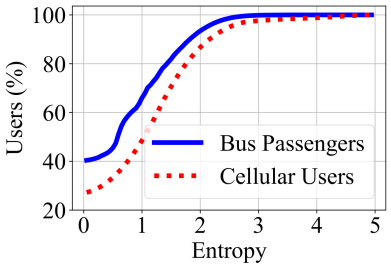


Fig. 5. Finger Trace Entropy

5 METHODOLOGY

In this section, we formalize our problem as a crowd uniqueness analysis task in bus WiFi systems. The *bus WiFi user uniqueness* is defined as the possibility that one user can be uniquely re-identified in a bus WiFi system. To study this problem, we design a model to analyze the crowd uniqueness from our large-scale dataset.

5.1 Terminology and Problem Definition

We summarize notations used in Table 2.

Table 2. Terminology and Notations

Terminology	Notation	Meaning
User	u_i	a bus WiFi user, u_i is the i_{th} user
Trace Record	r_i	a trace record in the bus WiFi system, r_i is the record from user i
Record Set	$\mathcal{R}_i$	trace record set generated by user u_i
Leaked Information	$\mathcal{L}_i^n$	set of n leaked records of user u_i
User Set	U_j	user set associated with record r_j
Identified User Set	U^n	users uniquely re-identified by n number of random records

Traces. A trace in bus WiFi systems is an observation of a bus WiFi user at a specific time, which are described by a tuple of attributes. In our analysis, there are two types of traces based on different human behaviors, i.e., (i) foot traces, which describe locations of users, (ii) finger traces, which describe the connection behavior of users. We use r_i to present a trace from user i. For different types, we use r_i^{finger} to present a finger trace and r_i^{foot} to present a foot trace.

Problem Definition. For each user u_i, we randomly select n number of records as a leaked dataset L_i^n. Therefore, given a uniqueness level n, we construct a leaked dataset $L_\forall^n$ for all users. Our goal is to find the number of users that can be uniquely re-identified with leaked data from the original dataset in the system, which is formalized in Equation 2 where p^n is the probability that a user can be uniquely re-identified by n random records.

$$U^n = \{u_i | L_i^n \cap R_\forall^n = u_i; i = 1, 2, ...\}$$
$$p^n = \frac{|U^n|}{|U|} \tag{2}$$

5.2 Analysis Model

The uniqueness analysis in such a large dataset is time-consuming and requires high memory cost. A straightforward method to compute uniqueness is based on two loops. We iterate one user from the leaked data in the outer loop and search for the matched users in the inner loop. The time complexity is $|\mathcal{L}| \cdot |\mathcal{R}|$ since we need a pair-wise check for all records between leaked data and original data. To reduce the computing complexity, we design a computing model named *PB-FIND* to calculate the uniqueness score given a uniqueness level n, which is the number of leaked records for every user. We illustrate our model in Figure 6, which includes three parts: (i) heuristic caching or indexing; (ii) candidate pruning; (iii) candidate set minimization.

(i) Heuristic Caching or Indexing. A user record in bus WiFi systems consists of two parts <u,r> where u is a user id and r is a studied trace record, e.g., URL and time in finger traces; location and time in foot traces; URL, location and time in hybrid traces. We use three maps in Table 3 to cache or index the relation between studied trace record r and user u in the bus WiFi dataset. To quantify the popularity of one trace r, we count the number of users associated with r and store their information in a hash map $\mathcal{M}_{score}$ of which the key is r and the value is the number of associated users. We store the mapping from r to the associated users in $\mathcal{M}_{candidates}$ of which the

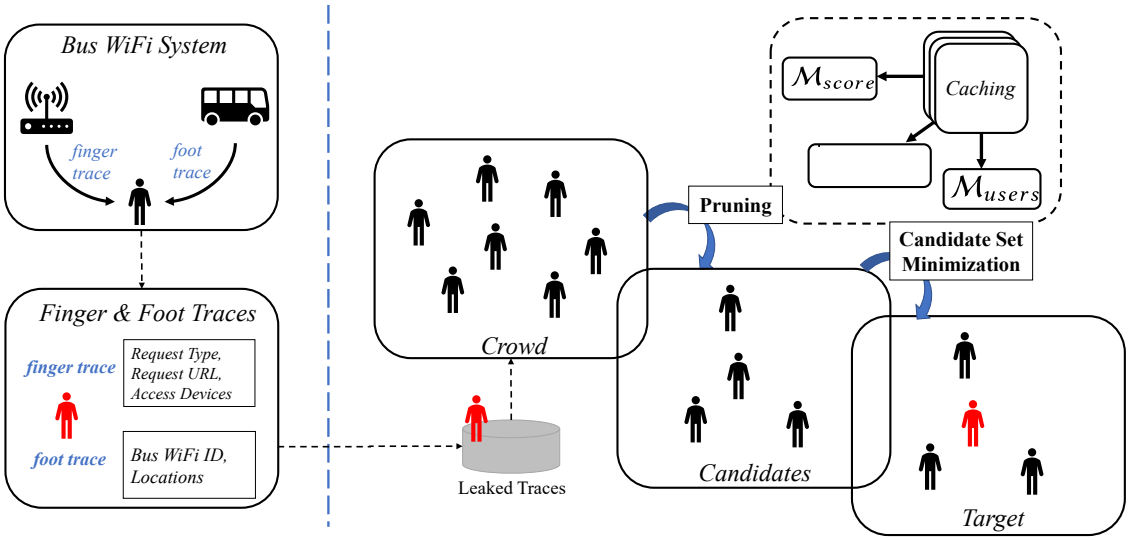


Fig. 6. PB-FIND Analysis Model

key is r and the value is a list of associated users. We cache number of <u,r> in $\mathcal{M}_{exists}$ of which the key is <u,r> and the value is the number of <u,r> in the original bus WiFi dataset.

Table 3. Heuristic Maps

Terminology	Notation	Meaning
heuristic score	$\mathcal{M}_{score}$	a hash map from traces r to number of users
candidate map	$\mathcal{M}_{candidates}$	a hash map from traces r to list of users
existence map	$\mathcal{M}_{exists}$	a true map indicating existence of <u,r> in row data

(ii) User Pruning. For one user and his leaked traces, we first find the trace with the smallest heuristic score in $\mathcal{M}_{score}$. In other words, we select the trace r_j that re-identifies the least number of users. Second, we use the candidate map $\mathcal{M}_{candidates}$ to retrieve all users with the trace. The retrieved users are candidates in our analysis. Our original problem is to re-identify a specific user from all users, of which the search space is the total number of users $|U|$. With the pruning process, the problem is transferred to re-identify a specific user from the candidate set U_j, of which the search space is significantly reduced to $|U_j|$.
(iii) Candidate Set Minimization. We start from a candidate set U_j and conduct a *candidate set minimization* process to filter unmatched candidates. The hash key of the existence map $\mathcal{M}_{exists}$ is the combination of a trace record and a user <u,r>; the value is the number of existence of the combination. We combine a user from the candidate set and a trace from the leaked user as a candidate key. We examine when the candidate user has the trace with existence map $\mathcal{M}_{exists}$. If the combination is not found in the trace map, we remove the user from the candidate set. We stop the process when either the size of candidate set is 1 or all leaked traces have been checked. In the first case, it means the target user can be uniquely re-identified with the leaked traces. In the second case, it means the target user cannot be uniquely re-identified. The details are given in Algorithm 1 with pseudo-code.
Online v.s. Offline Analysis: The proposed method is used for both offline and online analysis. In the offline analysis, we build indexing and caching on $\mathcal{M}_{score}$, $\mathcal{M}_{candidates}$ and $\mathcal{M}_{exists}$ and calculate the uniqueness based on current users. In the online analysis, for a new user or an existing user with $\mathcal{L}_i^n$, we add the new records

ALGORITHM 1: PB-FIND

Input : leaked record set $\mathcal{L}^n$, original record set $\mathcal{R}$
Result: $\mathbf{p}^n$
Initialize $\mathcal{M}_{score}$, $\mathcal{M}_{candidate}$, $\mathcal{M}_{exists}$;
Sort $\mathcal{L}^n$ by *score* from $\mathcal{M}_{score}$;
$N \leftarrow |\mathbf{U}|$;
for $i \leftarrow 1$ **to** N **do**
  $\mathcal{L}_i^n \leftarrow$ the leaked records of user u_i ;
  $r_0 \leftarrow$ check $\mathcal{M}_{score}$ to find the record with the least score in $\mathcal{L}_i^n$;
  candidates $\leftarrow$ get associate users of r_0 by $\mathcal{M}_{candidates}$;
  $\mathcal{COUNT}(candidate, r) \leftarrow$ count number of every pair <candidate,r> where candidate $\in$ candidates and r $\in \mathcal{L}_i^n$;
  during the count;
  if *<candidate,r> not in* $\mathcal{M}_{exists}$ ***or*** $\mathcal{COUNT}(candidate, r) > \mathcal{M}_{exists}(candidate, r)$ **then**
    remove candidate from candidates
  end
  if *|candidates| == 1* **then**
    number_of_unique_users += 1;
    break count;
  end
  end count;
end
$\mathbf{p}^n$= number_of_unique_users / N

with the user and re-sample n records $\hat{\mathcal{L}}_i^n$ for the user where n is the uniqueness level. Further, we update the three maps and update the uniqueness status of all other users with records associated $\Delta\mathcal{L}_i^n = \hat{\mathcal{L}}_i^n - \mathcal{L}_i^n$ and $\Delta\hat{\mathcal{L}}_i^n = \mathcal{L}_i^n - \mathcal{L}_i^n$.

6 RESULT & ANALYSIS

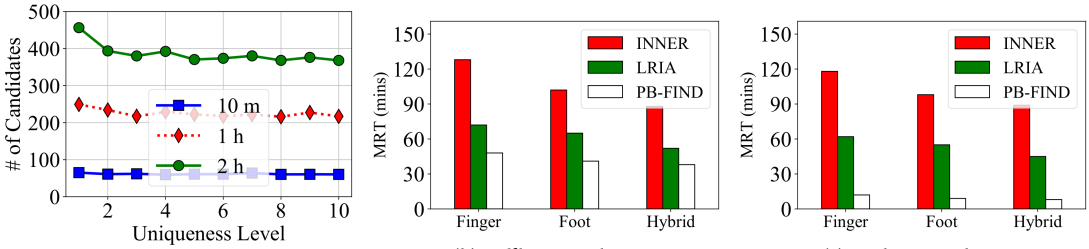
In this section, we evaluate our method and study the uniqueness among users.

6.1 Evaluation

6.1.1 Analysis. The time complexity in the map initialization process is $O(|\mathcal{R}|)$ since we loop all records to create three maps. In the pruning process, the algorithm iterates the number of leaked records of every user to find the record with the least score. The time complexity in this process is $O(|\mathbf{U}|)$, i.e.,$n \cdot |\mathbf{U}|$ where n is a small constant number, which is the uniqueness level. In the candidate set minimization process, the worst case is to check every pair of <candidate, record> for every leaked user. Therefore, the time complexity is $n \cdot \beta \cdot |\mathbf{U}|$ where β is the number of candidates. We investigate the average number of candidates with different n in Section 6. We found the number of candidates is relatively small. Therefore, we have $|\mathcal{R}| \geq n \cdot \beta \cdot |\mathbf{U}| > n \cdot |\mathbf{U}|$, the upper bound of the whole algorithm is a linear complexity $O(|\mathcal{R}|)$.

6.1.2 Settings. We introduce the evaluation settings for *PB-FIND* as follows. *(i) metrics:* we use *MRT* (Mean Re-identification Time) defined as $\frac{\sum_{i=1}^{n} t_i}{n}$, in which we calculate the average computing time to re-identify all users in the bus systems for n times of releases, as the metric to measure the computing efficiency in the re-identification. *(ii) baselines*: we compare our model with three baseline models, i.e., *Search*, *Inner Join* and *LRIA*. For a target user, in *Search*, we check records of every user in the system, the process will stop if there is a user with the

same records as the target user (not unique) or no such a user can be found (unique); The *Inner Join* is similar to *inner join* function in database SQL operation. We use *inner join* to join the release data with itself and the joined column is the released records. Specifically, it sorts released data and then uses two pointers (left pointer and right pointer) to scan the released data. If there is a match and the matched user is not the same user, both pointers move forward and the user on the left pointer will be marked as *not unique*. If two pointers are not matched, we move the smaller pointer forward. If the smaller pointer is the left pointer, we marked the user as *unique*. *LRIA* method is *Location Re-Identification Algorithm* proposed in [5] to find the unique PoIs that can be used to re-identify a surrounding user. We change the PoI to both finger traces and foot traces in our evaluation. *(iii) implementation:* we implemented both *PB-FIND* and baseline methods in finger trace, foot trace and hybrid trace uniqueness analysis in the bus WiFi system with more than 770 k users. We test the running time of our model with a desktop with 32GB memory, 1TB HDD storage, Intel Xeon CPU E5-1660 v3, installed with the latest Windows 10 and Python 2.7 coding environment. We will release the source code and the python package for the algorithms.



(a) Pruning (b) Offline Analysis (c) Online Analysis

Fig. 7. PB-FIND Evaluation

6.1.3 Evaluation Results. In the model analysis, we found the computing efficiency depends on the number of candidates after the pruning process. Figure 7a presents the number of candidates distribution after the pruning process. We found with the pruning process, the computing complexity decreases dramatically from 770 thousand bus WiFi users to hundreds of candidates on average. We further evaluate *PB-FIND* based on the proposed metric *MRT* in both offline analysis and online analysis. We found compared with other methods, *Search* method costs 10 times more computing time, therefore we omit the detailed comparison. As shown in Figure 7b, *PB-FIND* reduces the computing time around 50% compared with the two baseline methods. In the online analysis as in Figure 7c, *PB-FIND* reduces running time around 80% since it only needs to update the status of associate users.

6.2 Finger Traces

6.2.1 User Behaviors. To understand user behaviors in the bus WiFi system, we study their finger trace distribution based on the collected data. First, we found 26% of users only connect to one domain or web app and 56% of users connect to fewer than eight domains as shown in Figure 8. Besides, 74% users have less than 1-hour connection duration in Figure 9. The reason is that most buses cover short routes in small regions and the users rely on buses for short trips such as commuting between work and home locations. Figure 10 shows the traffic usage of bus WiFi users. We found two peaks in both downloading and uploading data, which is caused by different user behaviors. The first peak for two types of traffic is around 50 KB. These Users connect to bus WiFi in a short time period for online communication, e.g., they use WhatApp, WeChat or Skype for either personal or business purposes. The second downloading traffic peak is around 1 GB, and uploading traffic peak is around 20 MB. In this group, the users download or upload a large amount of data, e.g., searching for web pages, requesting online music or movies, posting on social media, etc.

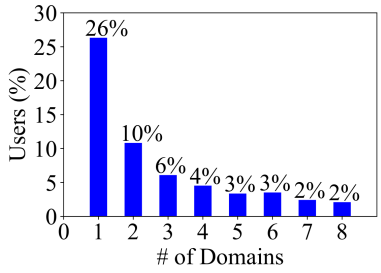


Fig. 8. Requests

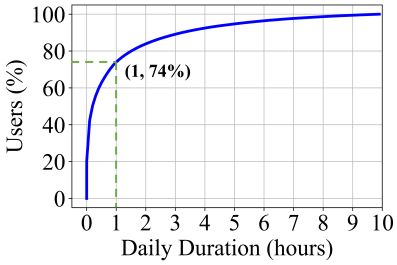


Fig. 9. Duration

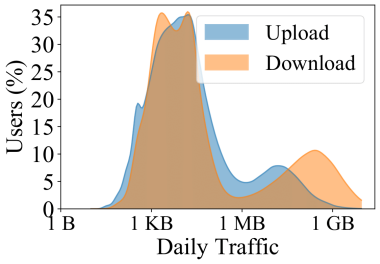


Fig. 10. Traffic

6.2.2 Finger Trace Uniqueness. To understand how unique users are with their finger traces, we analyze their connections to domains or web applications. We found an unbalanced distribution in domains. In particular, the users show a different Internet connection behavior compared with private networks such as cellular networks and home WiFi. For example, even though 10.8% of users visited TaoBao, which is the largest e-commercial web application in China, most of them only search for items or chat with sellers. Few users use financial services with bus WiFi connections, e.g., online payments such as Alipay, WeChat Pay, or PayPal. These are potentially caused by two reasons. First, buses are public environments and limited by space and functionality, e.g., a passenger is surrounded by other passengers when typing a payment password. Second, privacy is one of the major concerns for users due to the open nature of public WiFi.

The daily connection duration differs in visited domains as shown in Figure 11. The user behavior is impacted by the types of activities that they had. For example, most web browsing durations for online shopping website are less than 40 minutes, while the service website browsing durations are nearly uniformly distributed. In addition, users' PDF for different types of activity are likely to follow the Poisson distribution with respect to different parameters choose for population modeling. As as a result, it is not straightforward to re-identify a bus WiFi user with only historical visiting domains, e.g., *I saw user A visited google.com with bus WiFi*, without any temporal information, e.g., *I saw user A visited google.com last Friday*. As shown in Figure 12, only 20% of users can be uniquely re-identified by 10 visited domains; 40% of users can be uniquely re-identified by 20 visited domains. However, if a full visited URL is leaked for a specific user, e.g., leaked by service providers, hackers, malicious applications, a user can be easily re-identified from all bus WiFi users. 83.2% of users can be uniquely re-identified with less than 10 URLs and 91% of users can be uniquely re-identified with less than 20 URLs.

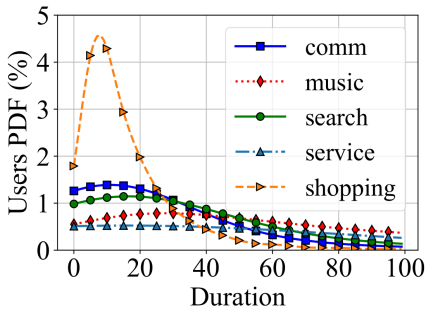


Fig. 11. Duration in Categories

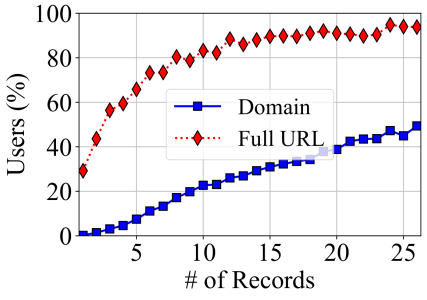
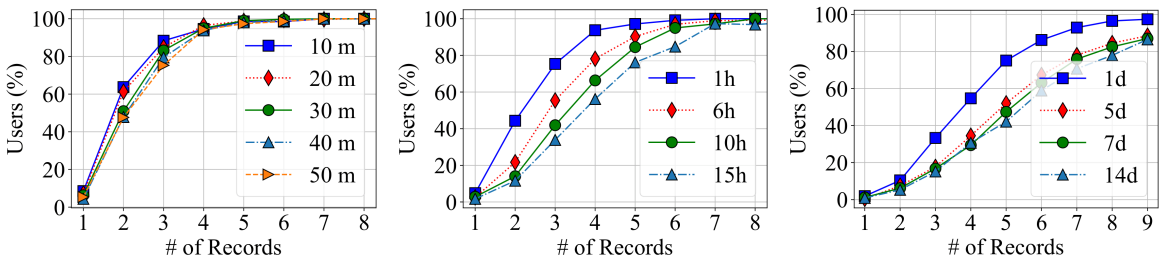


Fig. 12. Finger Trace Uniqueness

6.2.3 Impact of Temporal Granularity. We investigate how temporal granularity impacts the uniqueness of finger traces. A finger trace leakage with 1-hour temporal granularity means a finger trace is released with time and the

accuracy of associated time is at hour level, e.g., *I saw user A visited google.com around 5pm to 6pm last Friday.* We study the temporal granularity (i) within one hour from 10 minutes to 50 minutes in Figure 13a; (ii) within one day from 1 hour to 15 hours in Figure 13b; (iii) from 1 day to 2 weeks in Figure 13c. We found as the decrease of the temporal granularity, e.g., 10 minutes to 2 weeks, the uniqueness decreases. When the temporal granularity is 10 minutes, 97.8% of users can be uniquely re-identified with 4 random domain records. However, only 55.1% of users and 33.7% of users can be uniquely re-identified with 4 random records with 1-day and 1-week temporal granularity, respectively.



(a) Hour (b) Day (c) Week

Fig. 13. Impact of Temporal Granularity on Finger Traces

6.3 Foot Traces

Different from location records collected by location-based applications such as navigation apps, in which user actively report their locations for location-based services, foot traces in the bus system are passively sensed by the onboard GPS devices and uploaded by bus WiFi systems. Even though bus passenger behavior has been widely investigated in previous work [35] [25], most of them are based on pick-up locations from smart card transaction records or small samples of passengers [25]. Different from these works, bus WiFi systems collect detailed traces of users. Based on the collected foot traces, we analyze user behavior and their foot trace uniqueness to study the potential privacy risk.

6.3.1 Foot Trace Distribution. Figure 14 visualizes the heatmap of bus WiFi connections in four cities including a tier-1 city Shenzhen City, two provincial capital cities Nanjing and Changsha, a tier-2 city WuXi. We quantify the travel demand as the number of observed passengers. The red color indicates a higher demand and the green color indicates a lower demand. We found a higher demand in the tier-1 city due to a high coverage of the bus WiFi and a high traffic demand of the bus WiFi system. In general, we found the travel demand is positively related to the population density in the city as we compare travel demand distribution with population distribution from Worldpop dataset [26].

Figure 15 shows the number of buses a user takes in one day on average. We found most users have constant bus routes due to their fixed mobility patterns, e.g, commuting between several locations. 43.2% of users access to 1 bus WiFi device and 15.1% of users access to 2 bus WiFi devices during the one-day period. Buses are one of the major public transportation tools in cities, and most buses cover small regions. Therefore, most bus passengers are moving in regions with a certain radius. We first compute the maximum distance between two locations of a user in one day, which is defined as the daily radius at a specific day for the user. Second, we quantify the passengers' travel radius by the median value of the daily radius. We found the travel radius of 27% of bus WiFi users is smaller than 1 km and the travel radius of 91% of users is smaller than 10 km in Figure 16.

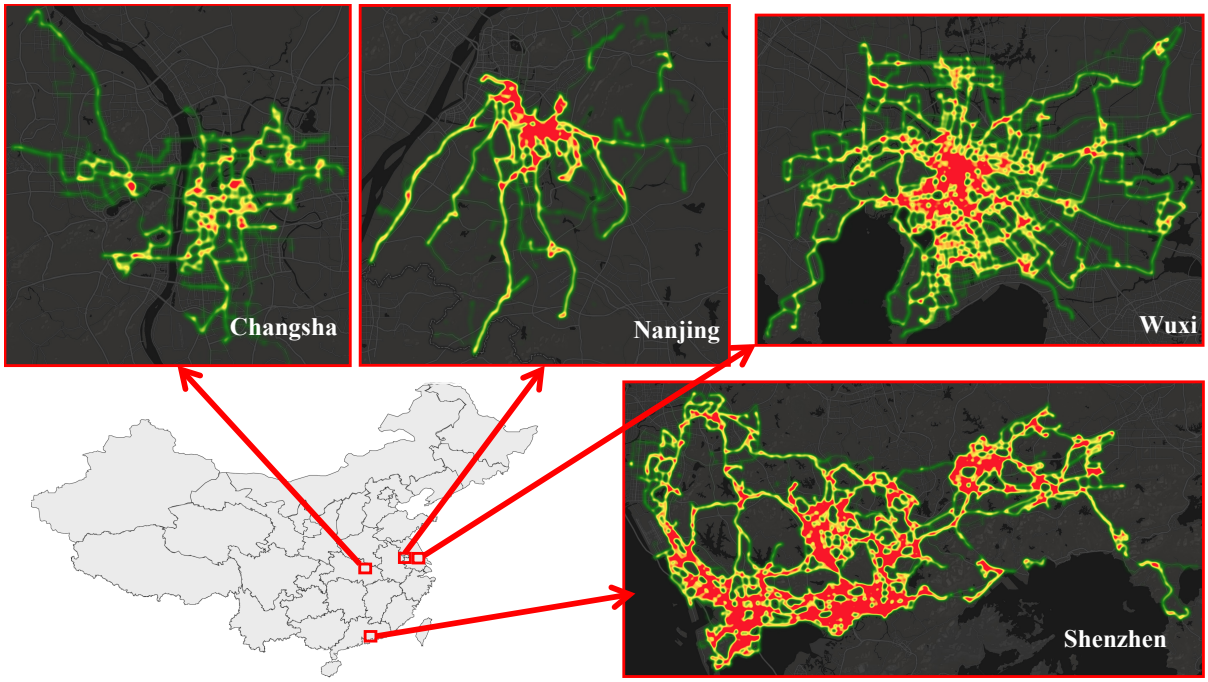


Fig. 14. Connection Activity distribution in Four Selected Cities

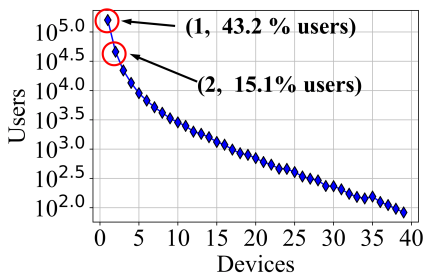


Fig. 15. Record

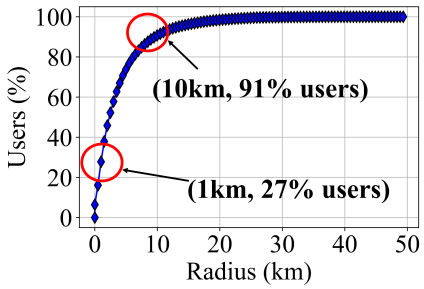


Fig. 16. Travel Radius

6.3.2 Foot Trace Uniqueness. We first study the uniqueness with bus WiFi devices in Figure 17, in which a user is re-identified with historical accessed WiFi devices, *e.g., I saw user A connected to the bus WiFi in bus 18.* We found it is with low probability to re-identify a bus WiFi user with only connected WiFi device leakage, e.g., 35.5% of users are uniquely re-identified with less than 10 historical connected bus WiFi devices, and 53.4% of users are re-identified with less than 30 historical bus WiFi devices accessed by a specific user. Figure 18 shows a comparison with respect to the user re-identification rate among different spatial granularity and the amount of spatial information being leaked, e.g., *I saw user A connected to bus WiFi near location l.* From the plot, while fewer than 3% of users can be re-identified with one grid information when the granularity is 1km, around 10% of them can be re-identified with 1 random record with a granularity of 1m. While around 40% of users can be re-identified by 10 locations with 1km granularity, 90% of users can be uniquely re-identified by same amount of

locations with 1m granularity. This indicates that a fine spatial granularity has a significant higher chance to expose an user's identity even when the spatial information provided is limited.

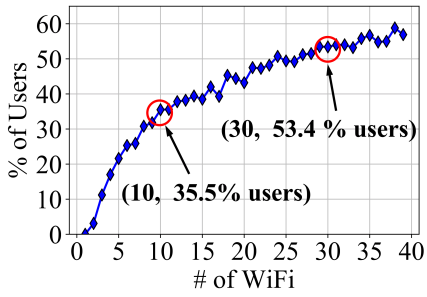


Fig. 17. Uniqueness With WiFi devices

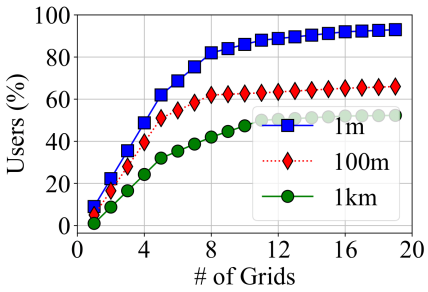


Fig. 18. Spatial Granularity

6.3.3 Impact of Temporal Granularity. We further investigate the privacy risk if the temporal information of connected devices are leaked, *e.g., I saw user A connected to the bus WiFi in bus 18 last Friday.* When combining temporal information with spatial information, we found even the most coarse spatial granularity (1 km) can uniquely distinguish over 65% of users by 4 records with a 10 minutes' precision as Figure 19 shows compared to Figure 18. On the other hand, the performance of unique user re-identification would not decrease significantly even with the most coarse temporal granularity (10 hours) compared with fine temporal granularity (10 mins) when the spatial information is given. The reason is that most users have daily patterns on buses,e.g., daily commuting between home and work locations, and when the temporal granularity is less than one day, the uniqueness will not change significantly. It further implies when spatial and temporal information are combined, even the most coarse granularity for both of them, it can uniquely re-identify more than 90% of users with more than 5 records.

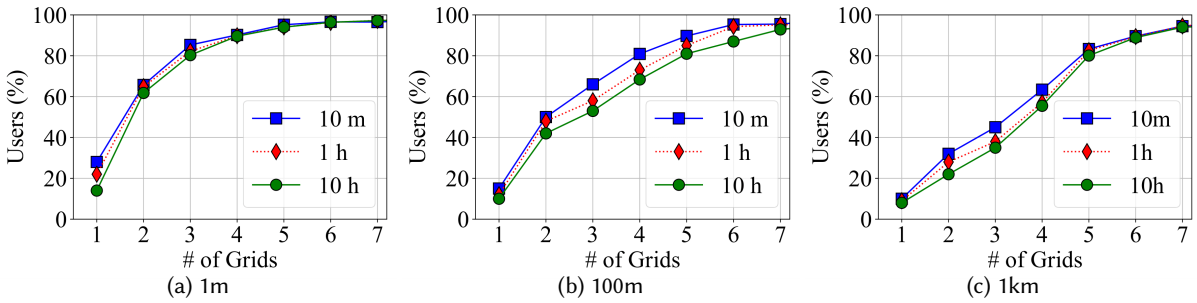
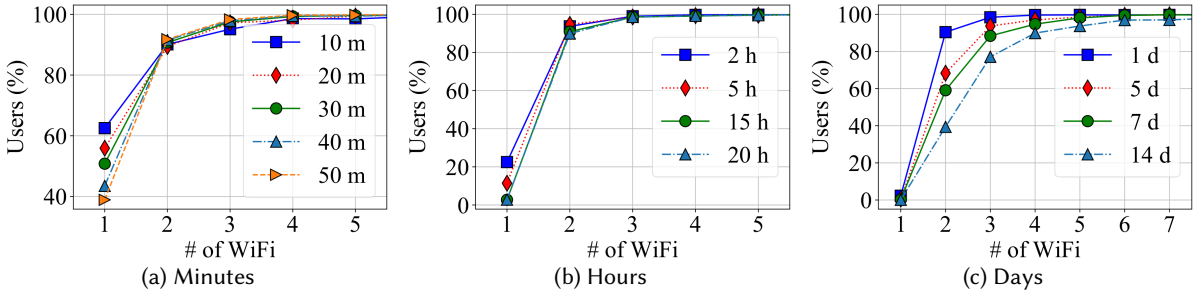


Fig. 19. Impact of Spatial and Temporal Granularity

According to Figure 20, leaking additional temporal information increases the uniqueness significantly. If 3 connected devices are leaked with temporal granularity less than one day, 97.2% of users can be uniquely re-identified from all bus WiFi users. The reason is when we combine temporal information and bus WiFi ID, the exact locations of users are determined. There is a high uniqueness on the combination of the three-dimension information, *e.g., I saw user A connected to the bus WiFi device in bus 18 at the airport station.* Since most passengers take one bus once in one day, the impact of temporal granularity on uniqueness is small when the temporal granularity is less than one day as shown in the three figures.



(a) Minutes (b) Hours (c) Days

Fig. 20. Impact of Temporal Granularity on Bus Traces

6.4 Hybrid Traces

6.4.1 Hybrid Trace Uniqueness. When both foot traces (foot trace can be represented by the bus WiFi ID and detailed locations) and finger traces are revealed, we found a much higher privacy risk to uniquely re-identify a user as shown in Figure 21, in which we use both domains and connected WiFi devices to identify a user, *e.g., I saw user A browsed google.com in bus 18.* We found 97.7% of users can be uniquely re-identified by hybrid traces with 5 random records even without temporal information.

6.4.2 Impact of Temporal Granularity. We study the impact of temporal granularity on the uniqueness of hybrid traces in Figure 22 and 23. When hybrid traces and temporal information are leaked, e.g., *I saw user B browsed google.com at location l around 4:10pm last Friday*, re-identification rate get much more increased. With a temporal granularity of 100m, more than 92% of the users can be re-identified with only one finger trace record. Compared to Figure 19, adding finger trace increases the uniqueness significantly from 61% to 92%. More than 99% of the users can be uniquely re-identified given 4 finger trace records even with a spatial granularity of 1km. This implies given that we get only 4 leaked browser records with rough information of the browsing time (morning, afternoon, evening) and the browsing location (within 1km grid), we are still very likely to uniquely re-identify this person.

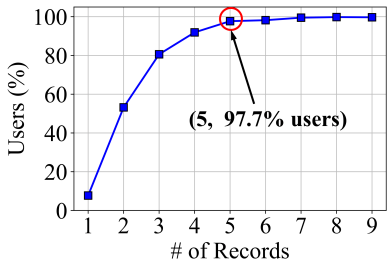


Fig. 21. Finger and Bus

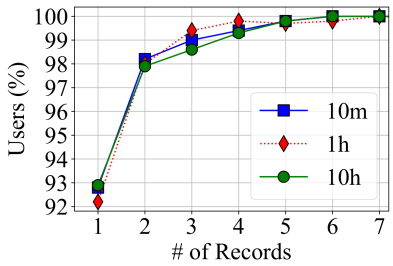


Fig. 22. Spatial and Temporal-100m

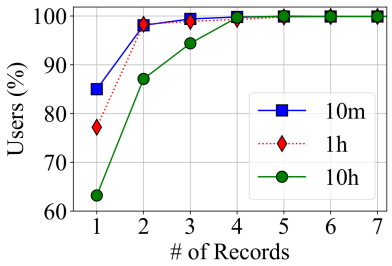
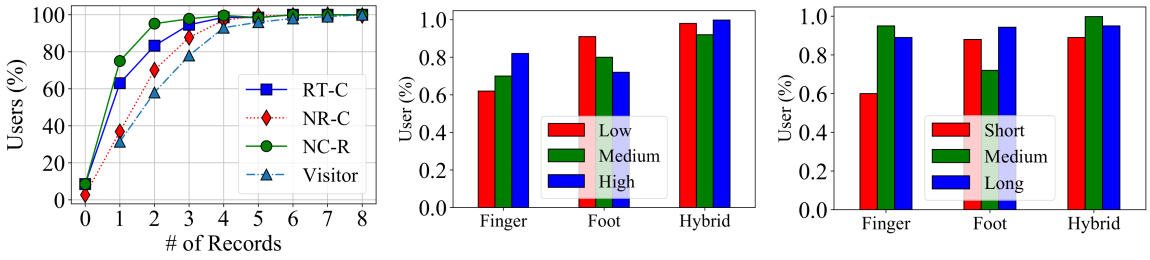


Fig. 23. Spatial and Temporal-1km

6.5 User Groups

We study bus WiFi privacy risk in four different user groups, e.g., routine commuters RT-C, non-routine commuters NR-C, non-commuting residents NC-R and visitors. To identify user roles in the system, we divide one day into four time slots, i.e., two daytime peak hours (7am-10am, 4pm-7pm), daytime non-peak hours (10am-4pm) and night time (8pm-7am). We first construct a daily vector for each user. A user is labeled as *commuter* in a day

when a user is observed in the two peak hours. We label a user as a *routine commuter*, i.e., *RT-C*, when a user is labeled as a *commuter* more than or equal to 3 days every week. A user is labeled as a *non-routine commuter*, i.e., *NR-C*, when a user is labeled as *commuter* less than 3 days every week. A user is labeled as a *non-commuting resident*, i.e., *NC-R*, if the user is connected to bus WiFi for more than 3 days every week but not a commuter. We label other users as *visitors*. We study the uniqueness of different groups of users as in Figure 24a. We found *NC-R* is the easiest to be re-identified from usage traces. The potential reason is that non-commuting residents visited random locations that are rarely explored by other users. Moreover, we study the impact of revisitation rate and different connection durations. We define the revisitation rate of one location as revisitation frequency on the location divided by the total number of visits. We first filter users with one record and the revisitation rate of a user is defined as the highest revisitation rate of all locations for the user. We study the impact of revisitation on the uniqueness of users with 4 released records in Figure 24b where the low revisitation rate is below 10%, the medium revisitation rate is between 10% to 30%, and the high revisitation rate is larger than 30%. We found that a higher revisitation rate makes users easier to be re-identified with finger traces. In contrast, users with a higher revisitation rate are less easy to be re-identified with foot traces and users with a medium revisitation are easier to be re-identified with hybrid traces. The potential reason is that a high revisitation rate decreases the diversity in foot traces. On the other hand, users with a high revisitation rate are mostly residents who connect to the bus WiFi system more frequently than non-resident users. Those users are easier to be re-identified by the finger traces. Moreover, we divide users into three groups based on their average connection duration, i.e., less than 5 minutes, between 5 minutes to 20 minutes, more than 20 minutes, and investigate the impact of the average connection duration on users in Figure 24c. We found users with medium connection duration is less easy to be re-identified by foot traces but easier to be re-identified by finger traces and hybrid traces.



(a) Uniqueness with User Groups (b) Revisitation Rate (c) Duration

Fig. 24. Uniqueness with Commuting Patterns

7 UNIQUENESS PROTECTION

In this section, we design a uniqueness protection model named *PB-HIDE*, which protects users from re-identification by potentially leaked information and preserves data utility by inserting a small amount of synthetic records in the original data.

As illustrated in Figure 25, if there are three users in the system, i.e., A, B, C. User A has two records, i.e., r_1 and r_2; user B has two records, i.e., r_1 and r_4; user C has two records, i.e., r_2 and r_4. If two records of user A, i.e., r_1, r_2, are leaked to an attacker. Without a protection strategy, user A can be uniquely re-identified from these three users. We are unable to change the leaked information since r_1 and r_2 are observed from real-world scenarios. To decrease the probability that a bus WiFi user can be uniquely re-identified from all records, one efficient strategy is to add synthetic records $\mathcal{N}$ as noise in the original records. For instance, when we add one record r_2 to user B or one record r_1 to user A in the original dataset $\mathcal{R}$, the user A is protected in the leakage. A simple approach is to add as many synthetic records as possible in the original data to decrease the privacy risk. However, this

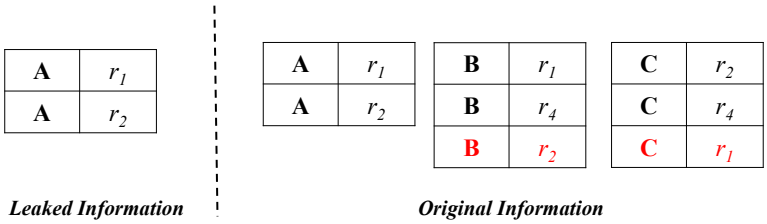


Fig. 25. Uniqueness Preserving with Redundant Records

method is not feasible in the bus WiFi system because the original data $\mathcal{R}$ are traded for social good or shared for various purposes, e.g., advertisers study user interests and customize advertisements to different user groups based on their finger traces and foot traces. The noise decreases the utility of the original dataset. Therefore, our target is to achieve the least uniqueness for privacy protection with a certain number of redundant records $\mathcal{N}$ in the original dataset $\mathcal{R}$.

Problem Definition: Given a potentially leaked dataset $\mathcal{L}^n$, the problem is to minimize the probability that users can be uniquely re-identified based on the potentially leaked data with m number of synthetic records. The optimization function is given in Equation 3 where $\mathcal{L}_i^n$ is n leaked records from user u_i and U^n is a set of identified users with n leaked records, m is the number of synthetic records.

$$\begin{aligned} &minimize\ p^n \\ &s.t.\ |\mathcal{N}| = m; \\ &p^n = |U^n|/|U|; \\ &U^n = \{u|\mathcal{L}_i^n \cap (\mathcal{R} \cup \mathcal{N}) = u_i\}; \end{aligned} \tag{3}$$

Methodology: To solve this problem, we design a greedy-based algorithm to insert synthetic records. First, we separate users into two groups, i.e., the identified group $\overline{U^n}$ and the unidentified group U^n. Since the users in the unidentified group are not uniquely re-identified, our algorithm is to reduce the number of users in the identified group until it researches the pre-defined threshold ϵ. We quantify the performance gain of inserting a record by the number of reduced identified users.

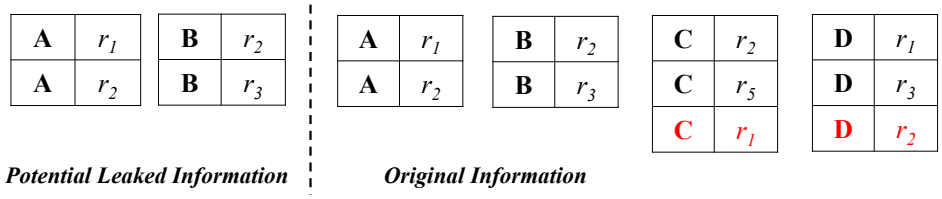


Fig. 26. Uniqueness Preserving with Redundant Records

As shown in Figure 26, both user A and user B can be uniquely re-identified based on their leaked records. We can add r_1 to user C and make user A not re-identifiable, the performance gain is 1. The second way is that we add one record r_2 to user D. As a result, both A and B are not re-identifiable. The performance gain is 2. The optimal solution starts from the record with the highest performance gain iteratively until the threshold is researched. However, quantifying the performance gain of every possible insertion, i.e., user and record pair <u,r>, requires a high computational cost. Instead, we solve the problem with a heuristic algorithm in two steps to

insert a synthetic record $< u, r >$: (i) *target user selection* to search for the target user u; (ii) *target record selection* to search for the r which achieves the highest performance gain in all possible combination of u.
(i) Target User Selection: In this process, we search for a user in the original dataset who potentially benefit the most among the identified users in the leaked dataset. Based on the example in Figure 26, we find user D is a better choice than C since it contains 2 records matched with the leaked dataset, i.e., r_1 and r_3, while user C only has 1 record in the leaked dataset. Therefore, we use a heuristic score s_i in Equation 4 that counts the number of common records in the user and re-identifiable users where $L^n_{U^n}$ is the leaked records from all re-identifiable users U^n.

$$\mathcal{I}^n_i = (\mathcal{R}_i \cap \mathcal{L}^n_{U^n})/\mathcal{L}^n_i; \;\; C^n_i = \mathcal{L}^n_{U^n}/\mathcal{I}^n_i; \;\; s_i = |\mathcal{I}^n_i| \tag{4}$$

(ii) Target Record Selection: In this step, we search for a record r_i that maximizes the performance gain when combined with user u_i. To find such a record r_i, we separate leaked records into two groups, i.e., $\mathcal{I}^n_i$ and C^n_i where C^n_i is a complementary set of $\mathcal{I}^n_i$ in terms of all leaked records $\mathcal{L}^n_{U^n}$ and $C^n_i = \mathcal{L}_{U^n}/\mathcal{I}^n_i$. The target record is selected from C^n_i. We calculate the number of co-existence between a record in C^n_i and all records in $\mathcal{I}^n_i$. In other words, every record in $\mathcal{I}^n_i$ has a vote. We select the record in C^n_i with the highest votes from $\mathcal{I}^n_i$. *PB-HIDE* integrates two steps in Algorithm 2.

ALGORITHM 2: PB-HIDE

Input : leaked record set $\mathcal{L}^n$, original record set $\mathcal{R}$,
Result: $\mathbf{p}^n$
$N \leftarrow |\mathbf{U}|$;
$k \leftarrow 0$;
while $k < m$ **do**
  $u \leftarrow$ apply target user selection ;
  $r \leftarrow$ apply target record selection ;
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{< u, r >\}$;
  $\mathcal{T} \leftarrow$ users not re-identifiable by adding <u,r> ;
  $U^n \leftarrow U^n/\mathcal{T}$;
  $k \leftarrow k + 1$
end

We use the example in the Figure 26 to illustrate the algorithm. In the first step, we calculate the heuristic score s_i for four users (A, B, C, D), $s_A = 1$, $s_B = 1$, $s_C = 1$, $s_D = 2$ since A has one record r_2 from leaked records except A itself, B has one record r_1 from leaked records except B itself, C has one record r_2 from leaked records, D has two records from the leaked records, i.e., r_1 and r_3. Therefore, we select D as the target user. In the second step, we split all leaked records $\mathcal{L}^2 = \{r_1, r_2, r_3, r_4\}$ to $\mathcal{I}^2_D = \{r_1, r_3\}$ and $C^2_D = \{r_2, r_4\}$. Therefore, the candidate records are r_3 and r_4. Since r_2 has two votes from A and B while r_4 has 0 vote, we select r_2 as the target record and insert <D,r_2> into the original dataset to protect the uniqueness privacy of both A and B.
Evaluation Settings: We introduce our evaluation settings as follows.

(1) Metrics: We use the uniqueness score, which is the percent of users that can be uniquely re-identified among all users, as our evaluation metrics.
(2) Baseline Approaches: We compare our algorithm with three baseline methods. *(i) RN:* RN stands for random noise, we select a random trace r from all records, a random user u, and insert <u,r> in the original dataset. *(ii) HMC:* is a state-of-the-art model for privacy protection in human traces. It infers human traces from sparse real locations and then generates synthetic data on traces [24]. *(iii) TN:* TN stands for target noise, similar to PB-HIDE, we select a trace r from leaked records, a random user u, and insert <u,r> in the original dataset.

(3) *Impact of Factors:* We study the impact of two factors. (i) *Impact of Trace Types:* we study the uniqueness change with synthetic records in different traces, i.e., finger traces, foot traces and hybrid traces. (ii) *Impact of Leaked Records:* we study the uniqueness change with a different number of leaked records, i.e., a different value of uniqueness level n.

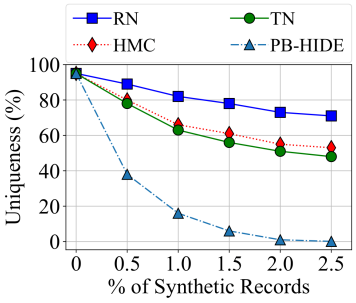


Fig. 27. Overall Performance

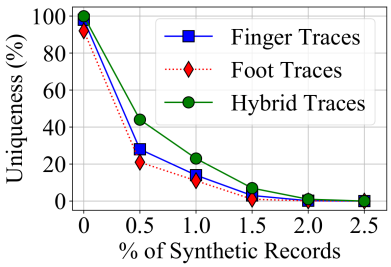


Fig. 28. Impact of Trace Types

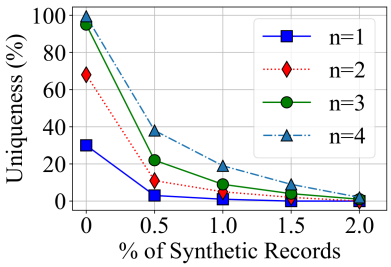


Fig. 29. Impact of Leaked Records

Evaluation Result: We implement the three models in a random leaked dataset $\mathcal{L}^4$ for all users. Compared with the three baseline approaches, *PB-HIDE* achieves the best performance in Figure 27. With 1.5% synthetic records, the uniqueness decreases to around 5% in *PB-HIDE* model while uniqueness is around 80% in *RN* and around 60% in *TN* and *HMC*. We further investigate the impact of two factors in Figure 28 and Figure 29. We found hybrid traces keep the highest uniqueness with the same amount of synthetic data. Figure 29 shows *PB-HIDE* achieves a decent performance with different leaked datasets. A small uniqueness level n, which is the number of leaked records per user, requires less amount of synthetic data to protect the same number of users.

8 DISCUSSIONS

Lessons learned: Based on the measurement results, we summarize a few lessons learned as follows.

(1) Our analysis results reveal bus WiFi users can be uniquely re-identified in bus WiFi systems with a very limited number of leaked records.
(2) Temporal information increases uniqueness significantly in both foot traces and finger traces. Specifically, with 4 random leaked traces, the uniqueness increases from 6% to 97.8% in finger traces (Figure 12 and Figure 13) and from 4.2% to 98% in foot traces (Figure 17 and Figure 20) by adding temporal information, e.g., 10-minute time slot. Therefore, sharing and storing data without temporal information will significantly decrease the privacy risk.
(3) Hybrid traces have a much higher uniqueness than either finger traces or foot traces (Figure 12, 17, 19, 21, 22, 23). Therefore, transferring and storing onboard sensor data separately can decrease the privacy risk in bus WiFi systems, e.g., transferring data with direct cellular connections instead of bus WiFi systems.
(4) With our protection model, we can protect most users from re-identification by potential leaked information with a small amount of synthetic records, e.g., reducing uniqueness from 98% to 3% with 1.5% synthetic records (Figure 27, 28, 29).

Limitation: First, we conduct the analyses and implement our system on the data from one bus WiFi company, Without the data access, we cannot validate our analysis result in other bus WiFi systems with a different number of users and deployment strategies. Second, our data does not cover all bus passengers in cities since (i) some buses are not upgraded with WiFi devices; (ii) some bus passengers prefer secure connections such as cellular networks. However, our analysis results provide several insights for privacy identification and protection in bus

WiFi systems. Moreover, we believe our analysis and protection techniques have the potential to be applied in many other systems, e.g., cellular systems, navigation systems, etc.

Implication. Since bus WiFi systems capture fine-grained finger traces and foot traces, it increases the risk that a user can be re-identified from the usage records. We found that a user can be re-identified more easily with hybrid traces compared with a single type of traces. Besides, the uniqueness is highly correlated with users' commuting patterns on buses. Based on our uniqueness analysis results, the bus WiFi company providing data access to us is building our protection model on the data which are shared with their collaborators such as advertisers. Since we found hybrid traces make a user much easier to be re-identified in bus WiFi systems, another interesting application is to recommend groups of users for Internet usage at certain locations or bus routes to protect user privacy. For example, most users in downtown areas with restaurants nearby are likely to use services such as Yelp, which makes a single user less easy to be re-identified.

Potential Societal Impacts: In the case study, we have revealed that there is a high probability that a user can be uniquely re-identified from bus WiFi systems via their foot traces and finger traces. Our study can be applied to more potential applications with better societal impacts. For instance, data holders can design better data sharing strategies for different purposes of data usage. Based on our analyses, we find separating multiple types of traces, decreasing spatial and temporal granularity, and incorporating synthetic data can decrease user uniqueness in the original dataset, which enable privacy-persevering data sharing mechanisms.

9 CONCLUSION

In conclusion, we design, implement and evaluate a privacy identification and protection system named *PrivateBus* based on a large-scale bus WiFi system with 770 thousand users, 20 million connection records and 78 million location records of individual users during a two-month period. We design two models, a uniqueness analysis model named *PB-FIND* to analyze the probability a user can be uniquely re-identified in the bus WiFi system with low computational cost and a uniqueness protection model named *PB-HIDE* to protect users from re-identification by potentially leaked information with a small amount of synthetic records. We divide user traces of bus WiFi systems into foot traces and finger traces, and study user uniqueness with different combinations of trace leakage. The measurement results reveal there is a high privacy risk in bus WiFi systems, e.g., 98.1% of users can be uniquely re-identified by only 2 random records if both of their connection records and locations are leaked to attackers. The evaluation results show *PB-HIDE* model protects 95% of users from potentially leaked information with 1.5% synthetic records added into the original dataset.

ACKNOWLEDGEMENT

This work is partially supported by NSF 1849238 and 1932223.

REFERENCES

[1] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. 2015. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 787–796.

[2] Giuseppe Ateniese, Randal Burns, Reza Curtmola, Joseph Herring, Lea Kissner, Zachary Peterson, and Dawn Song. 2007. Provable data possession at untrusted stores. In *Proceedings of the 14th ACM conference on Computer and communications security*. Acm, 598–609.

[3] Kwang-Hyun Baek, Sean W Smith, and David Kotz. 2004. A Survey of WPA and 802.11 i RSN Authentication Protocols. *Dartmouth Computer Science Technical Report2004* (2004).

[4] Hancheng Cao, Zhilong Chen, Fengli Xu, Yong Li, and Vassilis Kostakos. 2018. Revisitation in urban space vs. online: a comparison across pois, websites, and smartphone apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 156.

[5] Hancheng Cao, Jie Feng, Yong Li, and Vassilis Kostakos. 2018. Uniqueness in the city: Urban morphology and location privacy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 62.

[6] Deyan Chen and Hong Zhao. 2012. Data security and privacy protection issues in cloud computing. In *2012 International Conference on Computer Science and Electronics Engineering*, Vol. 1. IEEE, 647–651.

[7] Ningning Cheng, Xinlei Oscar Wang, Wei Cheng, Prasant Mohapatra, and Aruna Seneviratne. 2013. Characterizing privacy leakage of public wifi networks for users on travel. In *2013 Proceedings IEEE INFOCOM*. IEEE, 2769–2777.

[8] Elijah Chiland. [n.d.]. Free Wi-Fi service coming to 150 Metro buses around LA. https://la.curbed.com/2017/6/18/15827630/la-metro-wifi-bus-system-free-internet-los-angeles

[9] Chi-Yin Chow and Mohamed F Mokbel. 2009. Privacy in location-based services: a system architecture perspective. *Sigspatial Special* 1, 2 (2009), 23–27.

[10] VNI Cisco. 2018. Cisco Visual Networking Index: Forecast and Trends, 2017–2022. *White Paper* (2018).

[11] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.

[12] Marie Douriez, Harish Doraiswamy, Juliana Freire, and Cláudio T Silva. 2016. Anonymizing nyc taxi data: Does it matter?. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 140–148.

[13] Zhihan Fang, Wang Shuai, Guang Wang, Chaoji Zuo, Fan Zhang, and Desheng Zhang. 2020. CellRep: Usage Representativeness Modeling and Correction Based on Multiple City-Scale Cellular Networks. In *The World Wide Web Conference*. 1–11.

[14] Zhihan Fang, Yu Yang, Shuai Wang, Boyang Fu, Zixing Song, Fan Zhang, and Desheng Zhang. 2019. MAC: Measuring the Impacts of Anomalies on Travel Time of Multiple Transportation Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–24.

[15] Zhihan Fang and Desheng Zhang. 2017. Human mobility modeling on metropolitan scale based on multiple cellphone networks. In *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*. 321–322.

[16] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban population modeling based on multiple cellphone networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–25.

[17] Zhihan Fang, Fan Zhang, and Desheng Zhang. 2019. Fine-grained travel time sensing in heterogeneous mobile networks. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 420–421.

[18] Jon Fingas. [n.d.]. New York City rolls out its first WiFi-equipped buses. https://www.engadget.com/2016/05/17/new-york-city-wifi-buses-arrive/?guccounter=1

[19] Tiziano Inzerilli, Anna Maria Vegni, Alessandro Neri, and Roberto Cusani. 2008. A location-based vertical handover algorithm for limitation of the ping-pong effect. In *2008 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*. IEEE, 385–389.

[20] Simon L Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, and Vassilis Kostakos. 2015. Revisitation analysis of smartphone app use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1197–1208.

[21] Thivya Kandappu, Archan Misra, Shih-Fen Cheng, Randy Tandriansyah, and Hoong Chuin Lau. 2018. Obfuscation at-source: Privacy in context-aware mobile crowd-sourcing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 16.

[22] Arash Habibi Lashkari, Farnaz Towhidi, and Raheleh Sadat Hosseini. 2009. Wired equivalent privacy (WEP). In *2009 International Conference on Future Computer and Communication*. IEEE, 492–495.

[23] Hong Li, Limin Sun, Haojin Zhu, Xiang Lu, and Xiuzhen Cheng. 2014. Achieving privacy preservation in WiFi fingerprint-based localization. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2337–2345.

[24] Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. 2018. HMC: Robust Privacy Protection of Mobility Data against Multiple Re-Identification Attacks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 124.

[25] Hiroaki Nishiuchi, James King, and Tomoyuki Todoroki. 2013. Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. *International Journal of Intelligent Transportation Systems Research* 11, 1 (2013), 1–10.

[26] Population Reference Bureau. 2013. World Population Data Sheet. *http://www.prb.org/pdf13/2013-population-data-sheet-eng.pdf* (2013).

[27] Zhou Qin, Zhihan Fang, Yunhuai Liu, Chang Tan, Wei Chang, and Desheng Zhang. 2018. EXIMIUS: A measurement framework for explicit and implicit urban traffic sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 1–14.

[28] Muhammad Zubair Shafiq, Lusheng Ji, Alex X Liu, Jeffrey Pang, Shobha Venkataraman, and Jia Wang. 2013. A first look at cellular network performance during crowded events. *ACM SIGMETRICS Performance Evaluation Review* 41, 1 (2013), 17–28.

[29] Choonsung Shin, Jin-Hyuk Hong, and Anind K Dey. 2012. Understanding and prediction of mobile application usage for smart phones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 173–182.

[30] Kaixin Sui, Youjian Zhao, Dapeng Liu, Minghua Ma, Lei Xu, Li Zimu, and Dan Pei. 2016. Your trajectory privacy can be breached even if you walk in groups. In *Quality of Service (IWQoS), 2016 IEEE/ACM 24th International Symposium on*. IEEE, 1–6.

[31] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your apps give you away: distinguishing mobile users by their app usage fingerprints. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 138.

[32] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. 2017. Beyond k-anonymity: protect your trajectory from semantic attack. In *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[33] Guang Wang, Xiuyuan Chen, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. Experience: Understanding long-term evolving patterns of shared electric vehicle networks. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–12.

[34] Guang Wang, Wenzhong Li, Jun Zhang, Yingqiang Ge, Zuohui Fu, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. sharedCharging: Data-Driven Shared Charging for Large-Scale Heterogeneous Electric Vehicle Fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.

[35] Wei Wang, John P Attanucci, and Nigel HM Wilson. 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation* 14, 4 (2011), 7.

[36] Pascal Welke, Ionut Andone, Konrad Blaszkiewicz, and Alexander Markowetz. 2016. Differentiating smartphone users by app usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 519–523.

[37] Xiaoyang Xie, Yu Yang, Zhihan Fang, Guang Wang, Fan Zhang, Fan Zhang, Yunhuai Liu, and Desheng Zhang. 2018. coSense: Collaborative Urban-Scale Vehicle Sensing Based on Heterogeneous Fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–25.

[38] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. 2017. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1241–1250.

[39] Ye Xu, Mu Lin, Hong Lu, Giuseppe Cardone, Nicholas Lane, Zhenyu Chen, Andrew Campbell, and Tanzeem Choudhury. 2013. Preference, context and communities: a multi-faceted approach to predicting smartphone app usage patterns. In *Proceedings of the 2013 International Symposium on Wearable Computers*. ACM, 69–76.

[40] Yu Yang, Xiaoyang Xie, Zhihan Fang, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. VeMo: Enabling Transparent Vehicular Mobility Modeling at Individual Levels with Full Penetration. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[41] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 201–212.

[42] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K Dey. 2016. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 498–509.